REMARKS

Claims 1-6, 9-14, 17-27, 31, 35-44, 48, 52, and 53 are pending. Claims 1, 3, 9, 11, 17, 18, 35, 52, and 53 have been amended. Claims 2 and 10 have been canceled. No new matter has been entered. Claims 1, 3-6, 9, 11-14, 17-27, 31, 35-44, 48, 52, and 53 remain.

The amendments present the rejected claims in better form for consideration on appeal and may be admitted pursuant to 37 C.F.R. § 1.116(b)(2).

Claims 1-6, 9-14, 17-23, 35-40, 52, and 53 stand rejected under 35 U.S.C. § 103(a) as being obvious over International Application Publication No. WO 03/060766, to Lindh et al. ("Lindh"), in view of U.S. Patent No. 6,560,597, issued to Dhillon et al. ("Dhillon"). Applicant traverses.

Lindh teaches providing a source of synonymous words and expressions. A corpus of documents, stored in a database, are preprocessed, including performing word splitting, identifying proper names, removing stop words, applying a word stemming algorithm, and performing word weightings (Lindh, p. 19, lines 2-5). Following preprocessing, each unique term is assigned a weight according to that term's information content, which is determined using a Term Frequency times Inverse Document Frequency (TFIDF) (Lindh, p. 17, lines 21-23). Matrices are generated to describe relationships within the document corpus using the unique terms (Lindh, p. 18, lines 23-25). A document-concept matrix provides relationships between the documents in the corpus and concepts, which describe a particular document (Lindh, p. 19, lines 12-21). A term-document matrix provides relationships between the documents and unique terms selected from the documents (Lindh, p. 19, lines 22-28). A term-concept matrix receives information from the document-concept matrix and the term-document matrix to generate weight values representing relationships between the terms and the concepts (Lindh, p. 19, lines 29-32). The term-document matrix and the term-concept matrix are then used to generate a term-term matrix for describing relationships between the unique terms (Lindh, p. 20, lines 22-32). The term-term matrix is used for retrieving information from the document corpus (Abstract).

Lindh further teaches enhancing the above-described relationships by filtering the document corpus (p. 27, lines 18-25). A reduction in the number of similar documents in the corpus precludes large quantities of similar documents from biasing the relationship measures, which is characterized as a flaw that can be reduced using document clustering, such as *k*-means clustering (p. 27, line 25-p. 28, line 5). A representative document vector is generated for each cluster found by a clustering algorithm, such as by calculating a cluster centroid as the mean of all document vectors in the cluster (p. 28, lines 8-23). The representative document vector is added to the cluster and all other documents that belong to the cluster are removed from the initial document corpus (p. 28, lines 8-23).

In contrast, Dhillon teaches defining document concept decomposition vectors that represent a document vector space (Abstract). Documents, which are received from a text repository (Col. 3, lines 53-55), are parsed to remove redundant words for determining word frequency counts (Col. 4, lines 13-19). A disjoint clustering of the documents is performed by dividing the document vector space into partitions (Col. 5, lines 9-23). An initial set of concept vectors is determined as a mean vector of all document vectors for each partition (Col. 5, lines 56-60). A new partitioning of the document vector space is formed after the determination of the initial concept vectors (Col. 6, lines 288-30). New concept vectors corresponding to the new partitions are then formed (Col. 6, lines 37-45). The iterative partitioning continues until stop criteria, such as a magnitude of change or a predetermined number of iterations are satisfied (Col. 6, lines 46-58). Once stopped, the document vectors are projected into a subspace (Col. 6, lines 59-65).

The Examination Guidelines for Determining Obviousness Under 35 U.S.C. 103 in View of the Supreme Court Decision in *KSR International Co. v. Teleflex Inc.*, 72 Fed. Reg. 57,526 (Oct. 10, 2007) ("*KSR* Guidelines"), effective October 10, 2007, control obviousness determinations and provide exemplary rationales, as incorporated in MPEP 2143. Rationale (G), which includes some

teaching, suggestion, or motivation in the prior art that would have led one of ordinary skill to modify the prior art reference or to combine prior art reference teachings to arrive at the claim invention, appears to have been applied. Three factual inquiries must be made.

First, a finding must be made that there was some teaching, suggestion, or motivation, either in the references themselves or in the knowledge generally available to one of ordinary skill in the art, to modify the reference or to combine reference teachings. MPEP 2143(G)(1). Lindh teaches organizing document information in matrices to solve the problem of expressing term-to-term relationships and providing enhanced feedback for an initial search phrase (Lindh, p. 3, lines 10-22). A term-term matrix describes term relationships in the document corpus, including synonymous terms and related expressions (Lindh, p. 4, lines 2-7). The term-term matrix is enhanced by removing similar documents in the document corpus, using a clustering algorithm. The clustering algorithm generates clusters of similar documents. Those clusters having documents with high similarity are identified and a new document corpus is generated based on the initial corpus (Lindh, p. 6, lines 25-26; p. 27, lines 20-25). As a result, each cluster is represented by only a reduced set of documents in the new corpus (Lindh, p. 6, lines 26-32; p. 27, lines 25-28). Thus, Lindh is focused on a quality of a search using term-to-term relationships.

In contrast, Dhillon teaches partitioning a vector space into a set of disjoint clusters and assigning a concept vector for each partition. The partitioning is iteratively performed until the concept vectors satisfy a stopping threshold (Dhillon, Col. 6, lines 46-58). Dhillon focuses on solving the problem of efficiently implementing clustering to identify similar text documents in a collection by reducing the processing of the documents, such as by reducing a size of the document matrices (Dhillon, Col. 7, lines 8-17). For example, document vectors can require a database with upwards of a 100kx100k matrix size (Dhillon, Col. 2, lines 25-28). Through the iterative partitioning, the matrix size is reduced (Dhillon, Col. 5, lines 35-43). The partitioning technique is effective at reducing

dimensionality and yielding results commensurate with an initial partitioning of the document vector space (Dhillon, Col. 8, lines 1-5). Thus, Dhillon is focused on improving an efficiency of a search by reducing the document processing involved.

One skilled in the art would not be motivated to combine the teaching of Lindh with the teachings of Dhillon. Lindh focuses on improving a quality of a search to locate terms synonymous with a search term, whereas Dhillon focuses on improving search efficiency by iterative partitioning of documents. Lindh uses clustering as a method to reduce the number of similar documents in a document corpus, and not for a representation of the similar documents. Instead, Dhillon uses clustering as a method to partition and display documents for use in a search. Accordingly, a teaching, suggestion, or motivation to combine Lindh and Dhillon has not been shown.

Further, the Lindh-Dhillon combination teaches iterative partitioning of documents, which results in a concept vector subspace of document vectors (Dhillon, Col. 3, line 64-Col. 4, line 2). During partitioning iterations, a closest concept vector is found for each document vector to create new partitions (Dhillon, Col. 6, lines 31-37). New concept vectors are determined for the new partitions after each partitioning iteration (Dhillon, Col. 6, lines 37-40). Partition quality is determined by an objective function that measures the combined coherence of all clusters in a partition (Dhillon, Col. 7, lines 37-41). Quality values of the new and previous partitions are compared to determine a change in quality, which is based on a change between a previous grouping of documents and a new grouping of documents using mean concept vectors (Dhillon, Col. 7, lines 64-67). The change is applied to a stop threshold, which is determined by a *predetermined static value*, such as a magnitude of change in a concept vector from one iteration to another or number of iterations (Dhillon, Col. 6, lines 48-51; Col. 7, lines 64-67). If the threshold is met, the partitioning iterations will stop; otherwise, the partitioning iterations continue. Thus, the combination teaches comparing mean concepts of a partition to determine a value for applying a

predetermined threshold, rather than dynamically determining a threshold for each cluster as a function of the similarities between the documents grouped into each cluster based on the center of the cluster and the scores assigned to each of the concepts.

Moreover, modifying the teachings of Dhillon to consider dynamic data would not be predicable, as Dhillon teaches a static threshold for determining a stopping point for partitioning iterations. A fixed threshold is not adaptable, and replacing the fixed threshold with a dynamic threshold requires implementing functionality that continually adapts the threshold. Dhillon neither teaches nor suggests allowing the threshold to be dynamically redefined.

Second, a finding that there was reasonable expectation of success must be made. MPEP 2143(G)(2). Claims 1, 9, 17, 18, 35, 52, and 53 have been read on a combination of Lindh and Dhillon, but how the combination would be reasonably expected to succeed has not been explained. "The mere fact that references can be combined or modified does not render the resultant combination obvious unless the results would have been predictable to one of ordinary skill in the art." MPEP 2143.01(III) (citing *KSR International Co. V. Teleflex Inc.*, 550 U.S. __, __, 82 USPQ2d 1385, 1396 (2007)).

Further, the Lindh-Dhillon combination fails to teach each and every element of the claims. Independent Claims 1, 9, 17, 18, 35, 52, and 53 have been amended. Claim 1 incorporates the limitations of now-canceled dependent Claim 2, and now recites a scoring module determining a score, which is assigned to at least one concept that has been extracted from a plurality of electronically-stored documents, wherein the score is calculated as a function of a summation of a frequency of occurrence of the at least one concept within at least one such document, a concept weight, a structural weight, and a corpus weight. Claim 9 has been amended to incorporate the limitations of now-canceled dependent Claim 10. Amended Claim 9 now recites determining a score, which is assigned to at least one concept that has been extracted from a plurality of electronically-stored documents, wherein the score is calculated as a function of a summation of a

frequency of occurrence of the at least one concept within at least one such document, a concept weight, a structural weight, and a corpus weight. Claim 17 has also been amended to incorporate the limitations consistent with Claim 1, as amended. Amended Claim 17 recites code for determining a score, which is assigned to at least one concept that has been extracted from a plurality of electronically-stored documents, wherein the score is calculated as a function of a summation of a frequency of occurrence of the at least one concept within at least one such document, a concept weight, a structural weight, and a corpus weight.

Claims 18, 35, 52, and 53 have been similarly amended. Claim 18 has been amended to include the limitations of now-canceled Claim 2. Amended Claim 18 recites a scoring evaluation module evaluating a score to be associated with the at least one concept as a function of a summation of the frequency, concept weight, structural weight, and corpus weight. Claim 35 has been amended to include the limitations of now-canceled Claim 10. Amended Claim 35 recites evaluating a score to be associated with the at least one concept as a function of a summation of the frequency, concept weight, structural weight, and corpus weight. Claims 52 and 53 have been amended to include the limitations consistent with Claim 18, as amended. Claim 52 recites code for evaluating a score to be associated with the at least one concept as a function of a summation of the frequency, concept weight, structural weight, and corpus weight. Claim 53 recites means for evaluating a score to be associated with the at least one concept as a function of a summation of the frequency, concept weight, structural weight, and corpus weight.

In contrast, Lindh teaches conducting a search for related or synonymous terms using a term-term matrix for a corpus of documents. Prior to generating the term-term matrix, the document corpus is preprocessed. During preprocessing, terms are extracted and weighed. The term weight is based on an equation that accepts values for a number of occurrences of the term in a document, a total number of terms in the document, a number of documents in which the term exists, a total number of documents in the document corpus, and a value for a

position of the term in the document (Lindh, p. 17, line 20-p. 18, line 14). Once determined, the term weights are normalized and stored in a term-document matrix (Lindh, p. 18, lines 15-19). A ratio of the number of occurrences of the term in the document and the total number of terms in the document is multiplied by a negative log of the number of documents in which the term exists divided by the total number of documents in the document corpus, which in turn is multiplied by the position value of the term (Lindh, p. 17, line 20-p. 18, line 14). Thus, Lindh teaches calculating a weight for each term within a document corpus based on ratios for a number of occurrences of a term in the document based on a total number of terms in the document and a number of documents in which the term exists based on a total number of documents in the document corpus, rather than determining a concept score as a function of a summation of a frequency of occurrence of at least one concept, a concept weight, a structural weight, and a corpus weight.

Further, Lindh teaches a term-concept matrix, which provide relationships between terms and concepts. To calculate the relationship, the term weights and concept weights, both previously determined, are summed (Lindh, p. 22, line 26-p. 23, line 18). Thus, Lindh teaches a weighted relationship value between a certain term and a certain concept, rather than a score assigned to a concept that is calculated as a function of a summation of a frequency of occurrence of the at least one concept within at least one such document, a concept weight, a structural weight, and a corpus weight, which are each determined based on the concept.

Amended Claim 1 also recites forming the score assigned to the at least one concept as a normalized score vector for each such document and determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector. Claim 9 recites forming the score assigned to the at least one concept as a normalized score vector for each such document and determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector. Claim 17 recites code for forming the score assigned to the at least one concept as a

normalized score vector for each such document and code for determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector.

Similarly, Claim 18 recites a vector module forming the score assigned to the at least one concept as a normalized score vector for each such document in the electronically-stored document set and a determination module determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector. Claim 35 recites forming the score assigned to the at least one concept as a normalized score vector for each such document in the electronically-stored document set and determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector. Claim 52 recites code for forming the score assigned to the at least one concept as a normalized score vector for each such document in the electronically-stored document set and code for determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector. Claim 53 recites means for forming the score assigned to the at least one concept as a normalized score vector for each such document in the electronically-stored document set and means for determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector.

In contrast, Lindh teaches determining a weight for each term within a document. Upon determination, the weight is normalized and stored as a vector in a term-document matrix (Lindh, p. 18, lines 15-20). The vectors of the term-document matrix are used with data from a term-concept matrix to generate a term-term matrix, which contains vectors that describe conceptual relationships between terms (Lindh, p. 20, lines 22-25). A search for one or more terms is conducted using the term-term matrix. For example, a user can select one or more search term for which related concepts are returned (Lindh, p. 30, lines 7-10). The user then selects one or more of the related concepts. Documents that concern both the search term and the selected related concepts are returned (Lindh, p. 30,

lines 10-15). The concept selected by a user can bias the set of documents or rearrange the set (Lindh, p. 29, line 31-p. 30, line 2). To locate the biased information, a document corpus is generated based on selected terms (Lindh, p. 30, lines 18-21). A relationship value for each document is calculated from a document conceptual distribution, as determined by the document-concept matrix, and an input bias conceptual distribution received from a user (Lindh, p. 30, lines 23-30). A sum of the document conceptual distribution and the input bias conceptual distribution is calculated over every concept (*Id.*). Documents that are related to both the document conceptual distribution and input bias conceptual distribution are returned (Lindh, p. 30, lines 10-15; p. 30, line 31-p. 31, line 3). The document conceptual distribution and the input bias conceptual distribution are considered over all concepts to identify biased information, instead of comparing similarity values for each document. Thus, Lindh teaches returning documents, as a result of a search, based on a relationship value that includes input bias conceptual distributions, rather than determining a similarity between a normalized score vector for each document as an inner product of each normalized score vector.

Amended Claim 1 also recites a selection submodule selecting a set of candidate seed documents selected from the plurality of documents, a seed document identification submodule identifying a set of seed documents by applying the similarity to each such candidate seed document and selecting those candidate seed documents that are sufficiently unique from other candidate seed documents as the seed documents, a non-seed document identification submodule identifying a plurality of non-seed documents, a comparison submodule determining the similarity between each non-seed document and a center of each cluster, and a clustering submodule grouping each such non-seed document into a cluster with a best fit, subject to a minimum fit. Claims 9, 17, 18, 35, 52, and 53 recite limitations consistent with Claim 1, as amended.

In contrast, Lindh teaches document clustering to reduce a number of similar documents in a document corpus to prevent relationship bias between

terms (Lindh, p. 27, lines 25-28). Clusters are identified by a clustering algorithm, such as a k-means algorithm (Lindh, p. 28, lines 3-11). A representative document vector, generated by the clustering algorithm for each cluster identified, is determined by calculating a cluster centroid as the mean of all document vectors in the cluster (Lindh, p. 28, lines 11-14). The calculated representative document vector is then added to the cluster (Lindh, p. 28, lines 14-16). After determining a representative document vector for each cluster, a new document corpus is produced, in which each cluster is represented by a cluster representative vector (Lindh, p. 28, lines 20-23). The clustering algorithm is applied to the complete document corpus (Lindh, p. 28, lines 9-11), instead of being applied to a select portion of the document corpus. Thus, Lindh teaches applying a clustering algorithm to a document corpus, rather than selecting a set of candidate seed documents from a plurality of documents.

Further, Lindh fails to teach identifying a set of seed documents from the set of candidate seed documents. As described above, Lindh applies a clustering algorithm, such as a k-means algorithm to a document corpus to remove documents that are similar. After the clusters have been identified, a representative document vector is determined and assigned to each cluster (Lindh, p. 28, lines 20-23). Next, the representative document vector is added to the cluster, and documents belonging to the cluster are removed except for the representative document vector (Lindh, p. 28, lines 16-24; FIGURE 9A). As the clustering algorithm is applied to the complete document corpus, a set of candidate seed documents are not selected, nor is a set of seed documents identified based on a similarity determined for each document. Thus, Lindh teaches applying a clustering algorithm to a document corpus to identify clusters of the documents, rather than identifying a set of seed documents by applying the similarity to each such candidate seed document in each category and selecting those candidate seed documents that are sufficiently unique as the seed documents.

Moreover, Lindh fails to teach assigning non-seed documents into a cluster with a best fit, subject to a minimum fit. Documents can be clustered using a clustering algorithm, such as *k*-means clustering (Lindh, p. 28, lines 9-11). A set of clusters containing similar documents will be produced (Lindh, p. 28, lines 6-7). Thus, each document will be clustered with similar documents based on a particular algorithm without applying further requirements, such as a minimum fit criterion. Applying a minimum fit criterion to the teachings of Lindh would change the clustering of the documents since each document must satisfy additional criteria. For example, a document that is similar to a cluster will be placed in that cluster according to the clustering algorithm in Lindh. However, if a minimum criteria was applied, that same document may not be placed into the cluster, even though the cluster is similar, if the similarity fails to meet a minimum similarity. Therefore, Lindh teaches assigning documents to similar clusters using a clustering algorithm, rather than grouping a non-seed document into a cluster with a best fit, subject to a minimum fit.

Finally, additional findings must be made based on *Graham* factual inquiries, as necessary, in view of the facts of the case under consideration, to explain a conclusion of obviousness. MPEP 2143(G)(A). No further *Graham* factual findings were made.

"If any of [the three] findings cannot be made, then this rationale cannot be used to support a conclusion that the claim would have been obvious to one of ordinary skill in the art." MPEP 2143(G). Therefore, lacking sufficient findings, the combination of Lindh and Dhillon fail to render independent Claims 1, 9, 17, 18, 35, 52, and 53 obvious.

Claims 3-6 are dependent on Claim 1 and are patentable for the above-stated reasons, and as further distinguished by the limitations therein. Claims 11-14 are dependent on Claim 9 and are patentable for the above-stated reasons, and as further distinguished by the limitations therein. Claims 19-23 are dependent on Claim 18 and are patentable for the above-stated reasons, and as further distinguished by the limitations therein. Claims 36-40 are dependent on Claim 35

and are patentable for the above-stated reasons, and as further distinguished by the limitations therein. Withdrawal of the rejection is requested.

Claims 24-27 and 41-44 stand rejected under 35 U.S.C. § 103(a) as being obvious over Lindh and Dhillon as applied to Claims 18 and 35 above, and further in view of U.S. Patent No. 6,675,159, issued to Lin et al. ("Lin"). Applicant traverses.

Claims 24-27 are dependent upon Claim 18 and are patentable for the reasons stated above, and as further distinguished by the limitations therein. Claims 41-44 are dependent upon Claim 35 and are patentable for the reasons stated above, and as further distinguished by the limitations therein. Withdrawal of the rejection is requested.

Claims 30, 31, 47, and 48 stand rejected under 35 U.S.C. § 103(a) as being obvious over Lindh and Dhillon, and further in view of Lin. Applicant traverses.

Claims 30 and 47 have been canceled. Claim 31 is dependent upon Claim 18 and is patentable for the reasons stated above, and as further distinguished by the limitations therein. Claim 48 is dependent upon Claim 35 and is patentable for the reasons stated above, and as further distinguished by the limitations therein. Withdrawal of the rejection is requested.

The prior art made of record and not relied upon has been reviewed by the applicant and is considered to be no more pertinent than the prior art references already applied.

Claims 1, 3-6, 9, 11-14, 17-27, 31, 35-44, 48, 52, and 53 are believed to be in condition for allowance. Entry of the foregoing amendments is requested. Reconsideration of the claims, withdrawal of the finality of the Office action, and a Notice of Allowance are earnestly solicited. Please contact the undersigned at (206) 381-3900 regarding any questions or concerns associated with the present matter.

Respectfully submitted,

Dated: January 25, 2008

By: _____
Krista A. Wittman, Esq.
Reg. No. 59,594

Cascadia Intellectual Property
500 Union Street, Ste 1005
Seattle, WA 98101

Telephone: (206) 381-3900
Facsimile: (206) 381-3999